心理科学进展 2020, Vol. 28, No. 9, 1462–1477
Advances in Psychological Science https://dx.doi.org/10.3724/SP.J.1042.2020.01462

·研究方法(Research Method)·

心理与教育测验中异常反应侦查新技术：变点分析法*

张龙飞　王晓雯　蔡　艳　涂冬波

(江西师范大学心理学院，南昌 330022)

摘　要　变点分析法(change point analysis, CPA)近些年才引入心理与教育测量学，相较于传统方法，CPA 不仅可以侦查异常作答被试，还能自动精确地定位变点位置，高效清洗作答数据。其原理在于：判断作答序列中是否存在可将该序列划分为具有不同统计学属性两部分的点(即变点)，并且需使用被试拟合统计量(person-fit statistic, PFS)来量化两个子序列之间的差异。未来可将单变点分析拓展至多变点，结合反应时等信息，构建非参数化指标以及将现有指标拓展至多级计分或多维测验，以提高 CPA 的适用广度及效力。

关键词　异常反应；变点分析法；累积和法；被试拟合统计量

分类号　B841

1　引言

在心理测验中我们常常可以看到这样的现象：某被试在测验初期认真作答，但到后期失去了答题的兴致，于是便随意作答、乱选一通。也会在教育测验中注意到这样的情况：某考生在作答过程中发觉剩余考试时间不足，而后急匆匆地答题，顾不上仔细审题，导致许多本来能答对的题目却都答错了。研究者将此类现象统称为异常反应(aberrant response)。心理与教育测验中经常会出现各式各样的异常反应，主要包括：热身效应(warm-up effect)、加速作答(speededness)、疲劳(tiredness)、注意力不集中(loss of concentration)和对题目的预了解(item preknowledge)等(Sinharay, 2017b)。以“对题目的预了解”的现象为例，如果被试在考试前已经获取了题目的信息，那么作答会更加得心应手(Zhang, 2014)。因此，异常反应的存在会“污染”作答数据，如果数据“受污染”程度严重，会使参数估计精度大受影响并降低测验效度(Shao, 2016)。以往调查表明，在心理与教育测验中存在异常反应的被试占据了相当一部分比例。例如，Meade (2016)发现测验当中存在粗心作答(careless response)行为的被试一般约占总人数的10%。如果在研究中直接使用存在异常反应的数据进行分析，势必会影响研究结论的可靠性和可推广度。当前对于异常反应的侦查方法主要有两种：第一种是人工逐一检查数据，但是这种方法需要测验管理者亲力亲为对数据逐个审查，较为费时费力，而且由于这种方法具有很大的主观性，所以侦查的准确性也存疑；第二种方法是利用统计学手段对数据进行快速侦测，这种方法效率很高，可由计算机程序独立执行，更具客观性，但是侦测的准确度依赖于统计学手段的合理性。因此，开发并完善有效的异常反应侦测统计学方法便具有重要的理论与实际意义。

异常反应数据的侦查是统计过程控制(statistical process control, SPC)中的一种，传统上 SPC 采用的是累积和法(cumulative summation, CUSUM)。CUSUM 通过构造被试拟合统计量(person-fit statistic, PFS)实现对异常反应的侦测。基于 CUSUM 的 PFS 通过依题目顺序将各题上观察与期望得分间的残差累积求和得到，当其超出一定临界值，则意味着失拟(Sinharay, 2016)，即认为存在异常反应。这种方法最大的优点在于可以输出图像，具有可视化的特性，对整个作答序列能有清晰、

收稿日期：2019-10-12
* 国家自然科学基金项目(31960186, 31760288, 31660278)资助。
通信作者：涂冬波, E-mail: tudongbo@aliyun.com

直观和全面的把握。但是当侦测任务需要确定变点(change point)位置时, 测验人员必须亲自检查根据被试的作答序列生成的图像以定位变点。然而, 心理与教育测验的数据动辄由成百上千名被试的作答构成, 传统的 CUSUM 方法因此显得捉襟见肘了。新一代的异常反应侦查方法——变点分析法(change point analysis, CPA; Page, 1954)应运而生, 它可以克服传统方法的弱势, 更适应于心理与教育测验的环境。

CPA 是目前 SPC 中较为流行的方法, 它可以检测由一系列随机变量构成的序列中是否存在一个或多个变点, 并确定变点的位置。变点在通俗意义上是指 “模型中的某个或某些量起突然变化之点” (陈希孺, 1991)。在变点前后, 随机变量赖以生成的模型本身或模型参数会发生改变(Sinharay, 2017b), 或者说发生了结构性的变化(structural change)。CPA 最早用于生物学领域, 其研究肇始于 Page (1954)在 Biometrika 上发表的一篇关于连续抽样检验的文章, 后来被广泛应用于医学(Aminikhanghahi & Cook, 2017; Kass-Hout et al., 2012; Nam, Aston, & Johansen, 2012)、环境气候(Abahous, Ronchail, Sifeddine, Kenny, & Bouchaou, 2018; Suhaila & Yusop, 2018; Yu & Ruggieri, 2019)、金融(Allen, McAleer, Powell, & Singh, 2018; Thies & Molnár, 2018; Ye, Liu, & Miao, 2012)、工业(Maleki, Bingham, & Zhang, 2016; Mortaji, Noorossana, & Bagherpour, 2015; Nigro, Pakzad, & Dorvash, 2014)等各个领域。而在近些年才引入心理与教育测量。

CPA 可以用于侦查心理与教育测验中的异常反应现象, 异常反应的被试在作答过程中, 会出现作答表现在某道题后发生转变的现象, 这就是测量学意义上的变点。CPA 的优势在于, 它不仅可以鉴别某被试是否存在异常反应, 还能检测变点的具体位置(Yu & Cheng, 2019)。因此, 在数据分析中, 此方法能使测验人员对被试的异常部分数据单独进行清理(Embretson & Reise, 2000; Shao, Li, & Cheng, 2016), 而无需将该被试的所有数据删除, 以降低异常反应的影响, 最大程度保留有效数据并提升参数估计精度(Hong & Cheng, 2018; Patton, Cheng, Hong, & Diao, 2019; Yu & Cheng, 2019)。

无论是传统的 CUSUM 还是新兴的 CPA, 都需要通过构造 PFS 的方式来实现侦查。在心理与教育测量领域, CPA 构造 PFS 主要依托项目反应理论(item response theory, IRT)。根据美国《教育和心理测验标准》(Standards for Educational and Psychological Testing; American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014)的 4.10 条规定:当 IRT 模型用于测验开发时, 应当提供关于模型是否拟合的证据。而 PFS 可以量化被试的得分模式与 IRT 模型的拟合程度(Bradlow & Weiss, 2001), 因此可作为《标准》所需证据的一部分。现有 PFS 指标可以分为两类:参数化的(parametric)和非参数化的(non-parametric)。本文将要讨论的 CUSUM 和 CPA 两种方法的 PFS 都是参数化的指标, 即基于 IRT 进行构造。具体使用方法是: 通过将 PFS 与其在某一显著性水平下的临界值进行比较, 以鉴别被试是否存在异常反应。

当前, CPA 的研究在心理与教育测量领域已经取得了一些进展。研究表明: CPA 既可用于非自适应测验(传统纸笔测验), 也可以用于自适应测验(如计算机自适应测验) (Sinharay, 2016)。Zhang (2014)首次将 CPA 引入教育测验, 在计算机化自适应测验(computerized adaptive testing, CAT)的环境下中侦测是否存在已遭泄露的题目。Shao 等人(2016)成功将基于似然比检验的 CPA 运用于检测被试加速作答行为, 以识别被试的能力值是否存在个体内(intraindividual)变化, 并找到变化的发生位置。Shao (2016)进一步将 CPA 拓展至热身效应(warm-up effect)的侦查。Sinharay (2016)归纳了 CPA 的三种 PFS 指标, 我们将在后文对这三种 PFS 进行阐述。并且, 他还将 CPA 用于探测被试对题目的预了解现象, 并讨论了 CPA 在具体应用中的各项细节问题(Sinharay, 2017a, 2017b, 2017c)。Lee 和 von Davier (2013)使用 CPA 技术在一项国际语言评估测试的历年平均分上检测出了异常的变动, 这可以为测验管理者提供测试改革的依据。

本文将首先介绍心理与教育测量中常见的异常反应及其管理模型, 然后详细综述以往研究者构造的基于 CPA 和 CUSUM 两种方法的 PFS 及其临界值的确定方法, 并阐述 CPA 和 CUSUM 的操作流程, 之后综合比较两种方法在异常反应侦查

中的特点、优劣及使用时的注意事项，最后对于该研究领域当前存在的问题进行分析并指明未来的研究方向。通过合理运用 CPA，心理与教育测量学工作者可以更严谨高效地处理作答数据，提高研究的质量，本文还在前人的研究基础之上提出一些创新的观点，帮助启发后续研究者的思路并推动 CPA 的研究进程。

2 异常反应模型

常见的异常反应类型主要包括热身效应、加速作答、疲劳、注意力不集中和对题目的预了解等。这些异常反应的出现会降低测验效度并随之影响研究结论的可靠性，应当通过一定的技术手段准确高效地识别它们，以尽可能减小异常反应对于测验的影响。本节主要以测验中最为常见的异常反应之一——加速作答(speededness)为例进行论述，着重介绍加速作答的管理模型。加速作答模型可方便地拓展到其它异常反应的建模中，如热身效应(Shao, 2016)和后期随机作答(Yu & Cheng, 2019)等。建模研究能使人们深入理解异常反应的内在机制(Shao et al., 2016)，这对侦测领域的意义在于：通过加深对异常反应机制的理解，有助于开发和完善异常反应侦查的新方法和新指标。

2.1 传统 IRT 模型

在介绍异常反应模型前，需要先了解传统的 IRT 模型。传统项目反应理论模型包括正态肩形模型(the normal ogive model)、Rasch 模型和 logistic 模型。目前学界使用得比较多的是后两种模型，为方便讲解，在此以两参数的 logistic 模型(2PL logistic model)为例进行介绍，模型可以表达为：

$$P_{ij}(\theta)=P(X_{ij}=1\mid\theta_i,a_j,b_j)=\frac{\exp[a_j(\theta_i-b_j)]}{1+\exp[a_j(\theta_i-b_j)]},\quad(1)$$

其中，X_{ij} 是被试 i 在 j 上的作答，θ_i 为被试 i 的能力参数，a_j 和 b_j 分别是题目的区分度和难度参数，$P_{ij}(\theta)$ 为被试 i 在题目 j 上答对的概率。

2.2 加速作答及其模型

加速作答(speededness; Evans & Reilly, 1972)是指发生在速度非待测量构念的限时测验的一种效应。被试在测验后期的某道题处发觉作答时间不够，迫于时间压力加快作答速度，导致其作答表现持续下降到测验结束。基于不同基本假设，可将现有加速作答模型分为三类：混合模型(mixture model)、组合模型(hybrid model)和渐变模型(gradual change model)。

2.2.1 混合模型

为减轻测验中加速作答效应对参数估计造成的影响，Bolt, Cohen 和 Wollack (2002)在混合 Rasch 模型(mixture Rasch model; Rost, 1990)的基础之上对加速作答实施建模，该模型将所有被试分成两个类别：加速和非加速，且每名被试只归属于其中一类。在每个类里，被试在各题上的答对概率都可写作 Rasch 模型(即式(1)中 $a_j=1$)的形式，而每道题在两个类别上分别具有不同难度参数。通过对题目的难度参数施加一系列约束，以实现对加速作答的管理。例如，测验初期题目(未受加速作答影响)在加速和非加速类上的难度设置为相等，而对于后期题目(受到加速作答影响)，加速类上的难度参数比非加速类的大，作为对加速作答的惩罚。该模型为：

$$P_{ij|g}^{*}=\frac{\exp(\theta_{ig}-b_{jg})}{1+\exp(\theta_{ig}-b_{jg})},\quad(2)$$

其中，g 表示类别，可取 1 或 2 (代表加速或非加速类)。θ_{ig} 为在类别 g 中的被试 i 的能力参数，b_{jg} 是题目 j 在类别 g 上的难度参数。$P_{ij|g}^{*}$ 为被试 i 在题目 j 上的答对概率。该模型假定加速类下所有被试的变点位置一致。

由于被试在测验后期的作答易受加速作答影响(Oshima, 1994)，并导致数据受污染。因此，Bolt 等(2002)采用混合 Rasch 模型来修正测验后期题目的参数估计。结果表明：该模型不仅可改善参数估计的精度，还能有效对被试实施分类。此后，学者们深入研究并陆续将该模型拓展为其它形式，如混合两参数 logistic 模型(mixture 2PL logistic model; Bolt, Mroch, & Kim, 2003)、两维混合两参数 logistic 模型(two-dimensional mixture 2PL logistic model; de Boeck, Cho, & Wilson, 2011)和混合层级模型(mixture hierarchical model; Wang & Xu, 2015)等。

2.2.2 组合模型

Yamamoto 和 Everson (1997)构建了两参数组合模型(2PL hybrid model)，用于数据拟合并提高加速作答影响下的参数估计精度。模型假设加速作答被试经过变点后，作答策略将会从深思熟虑

转变为随机猜测。模型如下所示：

$$P_{ij|g}^{*}=\left[\frac{\exp[a_j(\theta_i-b_j)]}{1+\exp[a_j(\theta_i-b_j)]}\right]^{1-I_{j|g}}*(r_{j|g})^{1-I_{j|g}},\quad(3)$$

其中，$\frac{\exp[a_j(\theta_i-b_j)]}{1+\exp[a_j(\theta_i-b_j)]}$为传统 2PL 模型。$g$ 表示类别，但与混合模型中只有两个类别不同，此模型根据变点位置分类：同一变点位置的被试归为一类。$I_{j|g}$ 是指示函数(indicator function)：当 $I_{j|g}=0$，表示类别 g 的被试在第 j 题上正常作答，答对概率为 $\frac{\exp[a_j(\theta_i-b_j)]}{1+\exp[a_j(\theta_i-b_j)]}$；当 $I_{j|g}=1$，表示类别 g 的被试在第 j 题上存在加速作答，猜对概率为 $r_{j|g}$，数值等于第 j 题选项数目的倒数。因此被试在变点之前的题目上正常作答，而在变点后所有题上转变为随机作答，各题的答对概率是固定数值，这是一项严格的假设。

Yamamoto 和 Everson (1997)的研究表明：相较于传统的 IRT 模型，2PL 组合模型能有效地提升被试和题目的参数估计精度。并且，基于该模型特性，Yu 和 Cheng (2019)在模拟研究中将其改为多级计分的形式，以生成由于不专心所致的后期随机作答的数据。

2.2.3　渐变模型

组合模型认为被试经过变点之后，各题的答对概率会变成固定数值。然而，此处介绍的渐变模型对变点后答对概率的改变持有更加灵活的认识——该模型假设被试在变点后各题上的答对概率将会逐渐下降。Wollack 和 Cohen (2004)在研究中首次建立了渐变模型，目的是生成加速作答数据。此后，Goegebeur, de Boeck, Wollack 和 Cohen (2008)成功实现了模型的数据拟合和参数估计。两参数渐变模型(2PL gradual change model; Suh, Cho, & Wollack, 2012)为：

$$P_{ij}^{*}=\frac{\exp[a_j(\theta_i-b_j)]}{1+\exp[a_j(\theta_i-b_j)]}*\min\left(1,\left[1-\left(\frac{j}{J}-\eta_i\right)\right]\right)^{\lambda_i},\quad(4)$$

其中，$\frac{\exp[a_j(\theta_i-b_j)]}{1+\exp[a_j(\theta_i-b_j)]}$是传统 2PL 模型。$J$ 是题目总数。η_i ($0\leqslant\eta\leqslant1$)描述被试 i 在测验中开始加速作答的位置，数值上等于被试在加速前完成的题目数量占总题数的比例。例如 $\eta_i=0.8$ 表示该被试从测验的 80%位置之后开始加速作答。λ_i 是加速率(speededness rate)参数，用于控制答对概率 P_{ij}^{*} 的下降速度。λ_i 越大，答对概率下降得越快。假设某测验共 100 题，两名被试能力 θ 相等，η 值都为 0.8，即两人都在测验后 20 题上存在加速作答，而 $\lambda_1=1$，$\lambda_2=3$。当两人在第 90 题上作答时，对于被试 1，答对概率

$$P_{1,90}^{*}=\frac{\exp[a_{90}(\theta_1-b_{90})]}{1+\exp[a_{90}(\theta_1-b_{90})]}*\min\left(1,\left[1-\left(\frac{90}{100}-0.8\right)\right]\right)^{1}=0.9*\frac{\exp[a_{90}(\theta_1-b_{90})]}{1+\exp[a_{90}(\theta_1-b_{90})]},$$

而对于被试 2，答对概率

$$P_{2,90}^{*}=\frac{\exp[a_{90}(\theta_2-b_{90})]}{1+\exp[a_{90}(\theta_2-b_{90})]}*\min\left(1,\left[1-\left(\frac{90}{100}-0.8\right)\right]\right)^{3}=0.73*\frac{\exp[a_{90}(\theta_2-b_{90})]}{1+\exp[a_{90}(\theta_2-b_{90})]}.$$

即对于两名能力相等且变点位置一致的被试而言，由于 λ 取值差异导致两人在同一题上的答对概率相去甚远：被试 1 在第 90 题上的答对概率是正常答对概率的 0.9，而被试 2 的答对概率是正常答对概率的 0.73。可见，λ 取值对于被试在加速作答部分的答对概率影响很大。在(4)式中，若 $\eta_i=1$ 或 $\lambda_i=0$，意味着被试 i 在测验中不存在加速作答，此时式子变成传统 2PL 模型。

Shao 等人(2016)在研究中使用 2PL 渐变模型来生成加速作答数据，随后，Shao (2016)在该模型的基础上略作改动，构建了热身效应(warm-up effect)的管理模型，公式为：

$$P_{ij}^{*}=\frac{\exp\left[a_j\left(\theta_i-b_j\right)\right]}{1+\exp\left[a_j\left(\theta_i-b_j\right)\right]}*\min\left(1,\left[1-\left(\delta_i-\frac{j}{J}\right)\right]\right)^{\lambda_i}.\quad(5)$$

热身效应是指发生在测验初期的一种效应，被试由于不熟悉测验内容或者紧张等原因导致测验初期的作答表现会低于他的实际水平。当被试熟悉测验后，其作答水平将会恢复正常并在此后保持稳定的发挥。式中，δ_i($0\leqslant\delta\leqslant1$)描述被试 i 在测验何处摆脱热身效应的影响，数值上等于测验初期存在热身效应的题目数量占总题数的比例。例如，$\delta_i=0.2$ 表示被试 i 在测验的 20%位置之后摆脱热身效应，开始正常作答。其余符号意义与(4)式基本一致，不再赘述。

2.3 异常反应模型简评

本节以心理与教育测验中最为常见异常反应之一——加速作答为代表，详细综述了异常反应的管理模型。加速作答各模型可以很方便地拓展到其它异常反应的建模中，如后期随机作答和热身效应等。通过对模型的剖析和认识，有助于加深对异常反应内部机制的理解，从而为异常反应侦测新方法开发夯实理论基础。就加速作答而言，研究者们基于三类假设，从不同角度建立了三种模型。

混合模型把所有被试划分成加速和非加速两个类别，通过对两类上的题目参数设置约束来实现对加速作答的建模：在变点之前，两类的参数一致；而在变点后，两类的参数相异。对于该模型而言，所有加速作答被试的变点位置都是一样的。组合模型假设加速作答被试在经过变点之后，其作答策略将从深思熟虑转变为随机猜测，这项假设十分严格。并且，该模型允许被试有不同的变点位置。而最后介绍的渐变模型则具有更加灵活的视角：它假设经过变点之后，被试的答对概率会逐渐地降低。从这一点上，渐变模型相较于其它两个模型而言会更加符合实际情形(Goegebeur et al., 2008; Suh et al., 2012)：如果被试在测验后期某题处发觉剩余时间不足，便会加快做题速度，并随着剩余时间越来越短，会越做越快，导致他在每道题上的认知加工耗时越来越少，答对概率也越来越低。在渐变模型中，有一个关键的被试参数 λ，它表示受效应影响答对概率的下降速度，在模型中作为指数而存在。它的取值对于答对率影响很大，λ 值越大，答对率下降得越快。各被试的 λ 取值不一，表明有些被试受效应影响程度大些，有些受影响程度小些，这也符合实际情形。我们不禁可以想到被试的能力应该会与 λ 取值存在相关，因为按照一般逻辑，能力强的被试相比于能力弱的在面对时间压力时应更为从容镇定，即使在测验剩余时长不多的情况下，也更能抵抗这种负面干扰，充分调动认知资源解决问题。然而，究竟实际情况是否与猜测一致，还需未来研究去验证。

3 基于累积和法的异常反应侦查

在心理与教育测量领域，CUSUM 相对 CPA 出现得更早，以往研究者提出了多种基于 CUSUM 的 PFS (Bradlow & Weiss, 2001; Bradlow, Weiss, & Cho, 1998; van Krimpen-Stoop & Meijer, 2000, 2001, 2002)，由于本节旨在阐述此方法的思路，因此在这里仅介绍最基本的 CUSUM 的 PFS 指标——基于题目平均加权残差(averaged weighted residual)的 PFS。所谓“残差”，是指被试在某题目上观察与期望得分(由 IRT 模型预测)之间的偏离程度(Yu & Cheng, 2019)。因此，CUSUM 的基本思想在于：按照题目顺序依次将被试的观察与期望得分的残差累加来构造 PFS，以检测被试是否存在异常反应。传统的 PFS 指标由被试整个作答序列通过一次计算得到，并未将题目呈现的顺序纳入考虑，这会导致序列某处的正(负)残差被另一处的负(正)残差弥补，从而降低了传统指标的检测效果。而 CUSUM 在 PFS 构建上结合了题目顺序的信息。当它的指标超过特定临界值时，即判断为异常反应。假设现有某次测验的作答数据，该测验为 0-1 计分，共包括 J 题，为便于阐述，此后的讨论中将略去被试的下标 i。

3.1 基于单侧统计量的题目平均加权残差的 PFS

van Krimpen-Stoop 和 Meijer (2000) 以及 Meijer (2002)定义了基于单侧(one-sided)统计量的题目平均加权残差的两种 PFS 指标，这里的“单侧”是指此类统计量考虑了被试的作答表现的变化方向：向上的变化意味着被试的作答水平变高；向下的变化意味着被试的作答水平变低。基于题目平均加权残差的两种单侧统计量——“向上”(upper)统计量和“向下” (lower)统计量的公式如下：

$$C_0^+ = 0; C_0^- = 0, \tag{6}$$

$$C_j^+ = \max\{0, T_j + C_{j-1}^+\}; C_j^- = \min\{0, T_j + C_{j-1}^-\}, \tag{7}$$

$$T_j = \frac{1}{J}[X_j - P(X_j = 1|\hat{\theta})]. \tag{8}$$

如(6)式所示，C_j^+ 和 C_j^- 的初始值都为 0，由公式(7)可知，C_j^+ 恒为非负数，C_j^- 恒为非正数。现定义 C_j^+ 和 C_j^- 两个 PFS 的临界值分别为 UB 和 LB。当 $C_j^+ \geqslant UB$ 或 $C_j^- \leqslant LB$ 时，判定被试出现了异常作答。在 0-1 计分测验中，当被试答错题目 j 时，T_j 为负数；当答对题目 j 时，T_j 为正。如果被试从某题开始一直答对(答错)，则会导致 T_j 一直为正数(负数)，因此统计量 C_j^+ (C_j^-)会一直增大(减小)，当超过临界值后，作答将被判为异常。因此，基于 CUSUM 的 PFS 在应用中倾向将突然出现的一段“连贯”作答序列 (即一段得分多数为 0 或多数为 1 的序列)诊断为异常。一般而言，这种情况意味着

被试作答模式在此处产生了突然的变化，这种变化可能是由于疲劳、加速作答、注意力不集中或预先了解试题等原因所致(Sinharay, 2017b)。并且，从上述介绍中也可得知：由于 CUSUM 考虑到了题目顺序信息并采用基于累积和的统计量实施侦查任务，因此也具备检测作答序列中可能出现的多种异常效应的能力。

3.2　基于双侧统计量的题目平均加权残差的 PFS

在单侧统计量的基础上，Tendeiro 和 Meijer (2012)提出了基于双侧(two-sided)统计量的题目平均加权残差的 PFS：

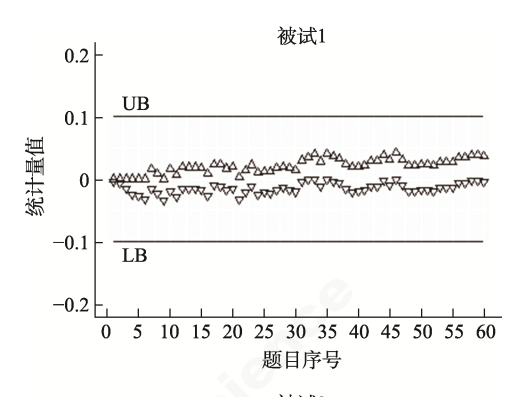
$$C^T = \max_{1\leqslant j\leqslant J} C_j^+ - \min_{1\leqslant j\leqslant J} C_j^-. \qquad (9)$$

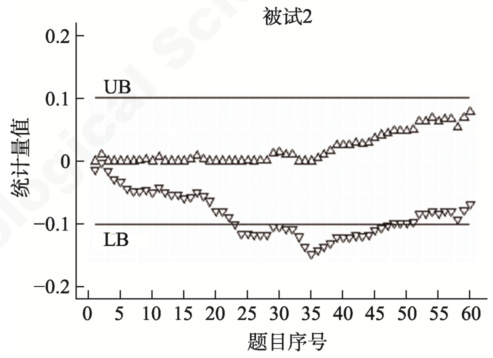
双侧统计量 C^T 整合了 C_j^+ 和 C_j^- 中的信息，其值等于整个作答序列中“向上”统计量 C_j^+ 和“向下”统计量 C_j^- 的最大差值。当 C^T 大于临界值时，判断该序列为异常。

除了上述基于题目平均加权残差($T_j = \frac{1}{J}[X_j - P(X_j = 1|\hat{\theta})]$)的 PFS，研究者还提出了其它 CUSUM 统计量，但是公式的表达形式都是一样的，只是将题目平均加权残差替换成其他内容，如对数似然比等(Armstrong & Shi, 2009; van Krimpen-Stoop & Meijer, 2001)，此处不再赘述。

3.3　CUSUM 图像的应用案例

为便于理解，此处以一个具体的 CUSUM 图像为例来介绍其使用方法和注意事项。图 1 展示了三名被试作答序列的 CUSUM 图像：被试 1 为正常作答的被试，被试 2 和被试 3 为异常作答被试。各图中正三角形表示 C_j^+，倒三角形表示 C_j^-，且中央两根水平实线分别代表 C_j^+ 和 C_j^- 的临界值，即 *UB* 和 *LB*。图中可以看出，两名被试的 PFS 都会在测验的某些位置超出临界值，因此两人的作答都被判定为异常。需要注意的是：CUSUM 的 PFS 是一种基于累积和的统计量，当被试在某题处的 PFS 超过了临界值，并不意味着该被试这题附近出现了异常。CUSUM 的 PFS 是不断累积计算的，应当取离达到临界值之前最近的 PFS = 0 的题目位置为变点估计值(Lai, 2001)。比如被试 3 的 C_j^- 虽然在第 53 题处低于 *LB*，但 PFS 从第 31 题处开始累积，说明他在该题附近开始出现异常作答。并且，被试 2 和被试 3 在测验的不同位置出现异常作答，根据异常出现的位置信息和具体形态可对其产生原因作初步推断，如被试 2 在测验初期大部分题目都答错了，导致出现了“向下”的异常，之后表现较好，原因可能是测验刚开始尚未熟悉测验内容。而被试 3 在测验后期出现的异常可能是因疲劳或加速作答所致。





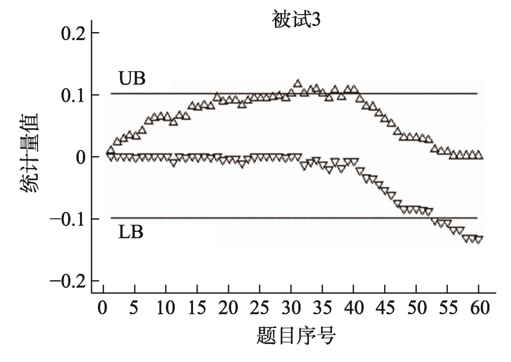


图 1　三名被试的 CUSUM 图像

3.4　两种 CUSUM 的 PFS 简评

CUSUM 是一种基于序列的技术，在每道题后立即更新统计量数值，由于可输出图像，该方法具有可视化的优势。通过观察图像，能迅速明晰异常反应发生的位置(Meijer, 2002)。凭借这种优势，CUSUM 还可在基于计算机(computer-based)

的测验中实施过程监控(process monitoring)：通过计算机程序实时监控被试的整个作答过程，当异常出现后迅速反馈给测验管理者，以便及时实施干预。但是在纸笔测验(P&P)中，则无法进行干预。

两种 CUSUM 的 PFS 中，基于单侧和双侧的指标各有所长。若在侦查伊始，测验管理者对于待侦查的异常反应类型有明确认识，建议选择基于单侧的 PFS。例如，当需要探测被试是否出现了加速作答，可直接选用“向下”的 PFS。而如果测验管理者的目标是侦查“笼统”而非某种特定的异常反应，即仅仅对于标定测验中的异常作答被试感兴趣，此时双侧 PFS 会比单侧 PFS 更适用。并且，实际测验中双侧统计量 C^T 往往会比单侧统计量 C_j^+ 和 C_j^- 更加有效，原因在于：实际测验中能力真值 θ_{True} 无法获知，所使用的是能力估计值 $\theta_{Estimate}$。假设被试 i 在测验的前半部分以真实能力 θ_{True} 作答，而在后半部分以更高(更低)的能力 $\theta_{Aberrant}$ 作答。此时根据整个作答序列估计得到被试能力值 $\theta_{Estimate}$，其取值必介于 θ_{True} 与 $\theta_{Aberrant}$ 之间。如此一来，单侧统计量的 PFS 在测验的两半部分(无论是正常还是异常部分)都会表现出异常：其中一半将会表现出“向上”的异常，另一半会表现出“向下”的异常。双侧 PFS 由于结合了两类单侧 PFS 的信息可有效避免这种情况发生(Armstrong & Shi, 2009)。通过基于 CUSUM 的 PFS 指标，可以清晰了解异常反应在作答序列中的位置。然而，当侦测任务需要确定变点位置时，测验管理人员须亲自检查 CUSUM 输出图像以定位变点，比较费时费力。接下来介绍的 CPA 可以免去人工检查的麻烦，由算法自动精准定位变点，有效节省人力资源。

4 基于 CPA 的异常反应侦查

CPA 以一种完全不同于 CUSUM 的视角来看待异常反应侦查的问题，它可以检测出由随机变量组成的序列中，是否存在一个或多个变点：在变点前后模型或模型参数是相异的。但与 CUSUM 一样，CPA 在使用中需构造 PFS 以实现侦测。

4.1 CPA 四种常用的 PFS

当前心理与教育测量学中常用的 4 种 CPA 的 PFS 主要有基于似然比检验的 $L_{\max}$，基于 Wald 检验的 $W_{\max}$，基于得分检验的 $S_{\max}$ 和基于加权残差的 $R_{\max}$。前 3 种 PFS 由 Shao 等人(2016)和 Sinharay (2016, 2017a, 2017b, 2017c)提出，而最后一种由 Yu 和 Cheng (2019)提出。4 种 PFS 构造的基本原理都在于：若某被试作答序列中存在变点，则此序列能以题目 n 为界划分为两个子序列：序列 1 为 X_1，$X_2,\cdots,X_n$，序列 2 为 $X_{n+1},X_{n+2},\cdots,X_J$。这两个子序列在某种统计学属性上具有根本性差异，CPA 的 PFS 可量化这种差异。若 PFS 在变点取 n 时数值最大且达到显著则表明第 n 题为变点位置。

4.1.1 基于似然比检验(likelihood ratio test)的 PFS

首先构造检验虚无假设 $\theta_{1n}=\theta_{2n}$ 并且基于似然比检验的统计量，当 n 取某已知的值时，对数化处理后的统计量如下所示：

$$\begin{aligned}L_n=-2\{&L(\hat{\theta}_0,\hat{\boldsymbol{\gamma}}_{\mathbf{0}};X_j,j=1,2,\cdots,J)-\\&L_{n1}(\hat{\theta}_{1n},\hat{\boldsymbol{\gamma}}_{\mathbf{1n}};X_j,j=1,2,\cdots,n)-\\&L_{n2}(\hat{\theta}_{2n},\hat{\boldsymbol{\gamma}}_{\mathbf{2n}};X_j,j=n+1,n+2,\cdots,J)\},\end{aligned}\tag{10}$$

其中，$\hat{\boldsymbol{\gamma}}$ 是由题目参数向量 $\hat{\mathbf{a}}$ 和 $\hat{\mathbf{b}}$ 构成的集合。$\hat{\theta}_0$ 和 $\hat{\boldsymbol{\gamma}}_{\mathbf{0}}$ 表示由该被试整体作答数据估计的参数，$\hat{\theta}_{1n},\hat{\boldsymbol{\gamma}}_{\mathbf{1n}}$ 和 $\hat{\theta}_{2n},\hat{\boldsymbol{\gamma}}_{\mathbf{2n}}$ 分别表示由变点前后作答序列所估计的参数。上述参数通过传统 IRT 模型估计得到。以 L_{n1} 为例，对数似然函数的具体展开形式为：

$$\begin{aligned}&L_{n1}(\hat{\theta}_{1n},\hat{\boldsymbol{\gamma}}_{\mathbf{1n}};X_j,j=1,2,\cdots,n)=\\&\sum_{j=1}^{n}[X_j\log P_j(\hat{\theta}_{1n})+(1-X_j)\log(1-P_j(\hat{\theta}_{1n})].\end{aligned}\tag{11}$$

为方便理解，此处只考虑一处变点的情况，且 $1\leqslant n\leqslant J-1$，在 n 取值范围内，定义基于似然比检验的 PFS 指标 L_{max} 为：

$$L_{\max}=\max_{1\leqslant n\leqslant J-1}L_n.\tag{12}$$

$L_{\max}$ 检验的虚无假设 H_0 为：此序列不存在变点(即对于取值范围内任意 n，$\theta_{1n}=\theta_{2n}$ 始终成立)；对应的备择假设 H_a 为：序列至少存在一处变点。对于 L_n 而言，在虚无假设下服从自由度为 1 的渐近 χ^2 分布，这是因为：似然比检验的统计量服从自由度为两嵌套模型未知参数个数之差的渐近 χ^2 分布，此处的约束模型(虚无假设模型)与无约束模型(备择假设模型)相比只多了一个约束条件($\theta_{1n}=\theta_{2n}$)，两模型中未知参数个数相差 1，因此这里的自由度为 1，后文将介绍的基于 Wald 检验的和得分检验的 PFS 同理。然而，因 L_{max} 取 L_n 的最大值，所以 L_{max} 不服从某个自由度已知的卡方

分布(Chen & Gupta, 2012), 并无建议的分布，它的虚无假设分布(null hypothesis distribution)能通过蒙特卡洛模拟获得，并可以根据虚无假设分布得到各显著性水平上的临界值 L_C。L_C 作为临界值，可用于判断基于似然比检验的 L_{max} 是否越界。若 $L_{max} > L_C$，则认为该作答序列存在变点，并得到变点的具体位置 n，即在题目 n 后被试的作答发生改变。若 $L_{max} < L_C$，则认为该被试作答正常，不存在变点。

L_{max} 适用于双侧检验(two-sided test)，以检验虚无假设 H_0 的正确性。当侦查目标仅仅是检测被试的作答序列中是否存在变点而不考虑能力变化方向时，采用这个统计量是合适的。然而，当侦查任务是检验由某种目标效应(如加速作答)导致产生的变点时，则需对 L_{max} 进行变换。例如，当检测某个可能存在加速作答的作答序列时，检验的虚无假设 H_0 为：对于取值范围内任意 n，$\theta_{1n} \le \theta_{2n}$ 始终成立。这种情况下，引出 L_{max} 的单侧检验形式(Sinharay, 2017a)，首先有：

$$L_{sn} = \begin{cases} \sqrt{L_n}, 若 \hat{\theta}_s \geqslant \hat{\theta}_{\bar{s}}, \\ -\sqrt{L_n}, 若 \hat{\theta}_s < \hat{\theta}_{\bar{s}}. \end{cases} \tag{13}$$

其中，$\hat{\theta}_s$ 是基于受某效应影响的作答序列估计能力值，$\hat{\theta}_{\bar{s}}$ 是正常作答序列估计能力值。L_{sn} 是单侧检验统计量，其绝对值等于 L_n 的平方根，并且在虚无假设下服从渐近的标准正态分布，原因在于：在虚无假设下 $\hat{\theta}_s > \hat{\theta}_{\bar{s}}$ 与 $\hat{\theta}_s < \hat{\theta}_{\bar{s}}$ 是等可能出现的，因此 L_s 取正或负的符号次数也是趋近的，且由于 L_{sn} 的绝对值等于 L_n 的平方根，L_n 服从自由度为 1 的渐近 χ^2 分布，所以 L_{sn} 在虚无假设下服从渐近标准正态分布。当异常部分的能力估计值高于正常能力估计值时，使用正的统计量进行检验，否则使用负的统计量。

由此，单侧 PFS 指标 L_s 可表达为：

$$L_s = \begin{cases} = \max\limits_{1 \leqslant n \leqslant J-1} L_{sn}, H_0: 对任意n都有 \theta_s \geqslant \theta_{\bar{s}}, \\ = \min\limits_{1 \leqslant n \leqslant J-1} L_{sn}, H_0: 对任意n都有 \theta_s \leqslant \theta_{\bar{s}}. \end{cases} \tag{14}$$

例如，当需要检验某被试是否存在加速作答时，使用的 PFS 指标为 $L_s = \min\limits_{1 \leqslant n \leqslant J-1} L_{sn}$，此时若 L_s 显著低于临界值，则拒绝虚无假设，认为被试存在加速作答行为并可由此定位变点。

4.1.2　基于 Wald 检验(Wald test)的 PFS

基于 Wald 检验的统计量也可用于检验虚无假设 $\theta_{1n} = \theta_{2n}$ 的正确性。当 n 取某给定值时，统计量的公式如下：

$$W_n = \frac{(\hat{\theta}_{1n} - \hat{\theta}_{2n})^2}{\dfrac{1}{I_{1n}(\hat{\theta}_0)} + \dfrac{1}{I_{2n}(\hat{\theta}_0)}}, \tag{15}$$

其中，I 表示对应作答序列(序列 1 和 2)所有题目的 Fisher 信息量总和。题目的信息量是 IRT 中用于衡量某题对特定能力值被试可提供测量精度的指标，信息量越大，表示该题对于这种能力被试的测量效果越好。注意：此处计算信息量使用的是通过整个作答序列估计的能力值 $\hat{\theta}_0$。

题目 j 的信息量公式为

$$I_{j(\theta)} = \frac{1.7^2 a_j^{\ 2}}{e^{1.7a_j(\theta - b_j)}[1 + e^{-1.7a_j(\theta - b_j)}]^2}, \tag{16}$$

与 L_{max} 类似地，W_{max} 表示为：

$$W_{max} = \max_{1 \leqslant n \leqslant J-1} W_n. \tag{17}$$

同样地，该指标检验的虚无假设 H_0 为：此序列不存在变点；备择假设 H_a 为：序列至少存在一处变点。Andrews (1993)以及 Csorgo 和 Horvath (1997)发现，当变点位于作答序列最前或者最后几题时，W_{max} 的侦测效力(power)将会变得十分小。因此，Andrews (1993)建议将 n 限定在整个作答序列的中间约 70%的范围，即 $W_{max} = \max\limits_{J_1 \leqslant n \leqslant J-J_1} W_n$，$J_1$ 取靠近 0.15J 的整数，以增强检测效力。对于前述 L_{max}，也可以在使用时做此限定，提高侦测效力。

基于 Wald 检验的统计量 W_{max} 适用于双侧检验，当进行单侧检验时，只需对(15)式右侧开根号，变成单侧检验统计量 W_{sn} (Estrella & Rodrigues, 2005)：

$$W_{sn} = \frac{\hat{\theta}_{1n} - \hat{\theta}_{2n}}{\sqrt{\dfrac{1}{I_{1n}(\hat{\theta}_0)} + \dfrac{1}{I_{2n}(\hat{\theta}_0)}}}, \tag{18}$$

此时，单侧检验的指标 W_s 表达式为：

$$W_s = \begin{cases} = \max\limits_{1 \leqslant n \leqslant J-1} W_{sn}, H_0: 对任意n都有 \theta_{1n} \leqslant \theta_{2n}, \\ = \min\limits_{1 \leqslant n \leqslant J-1} W_{sn}, H_0: 对任意n都有 \theta_{1n} \geqslant \theta_{2n}. \end{cases} \tag{19}$$

若 W_s 临界值的绝对值为 h，当 W_s 显著大于临界值 h 时，可拒绝虚无假设 H_0：对任意 n 有 $\theta_{1n} \leqslant \theta_{2n}$。即认为序列中存在变点且被试在变点前的能力高于变点后能力，例如在测验中出现加速作答；当 W_s 取值显著小于临界值$-h$ 时，则可拒绝虚无假设 H_0：对任意 n 有 $\theta_{1n} \geqslant \theta_{2n}$，即认为序

列中存在变点且被试在变点前能力低于变点后能力，例如出现热身效应。

4.1.3 基于得分检验(score test)的 PFS

基于得分检验的统计量 S_n 可检验虚无假设 $\theta_{1n}=\theta_{2n}$，当 n 取某给定值时，表达式为：

$$S_n=\frac{[\nabla(\hat{\theta}_0;X_j,j=1,2,\cdots,n)]^2}{I_{1n}(\hat{\theta}_0)}+\frac{[\nabla(\hat{\theta}_0;X_j,j=n+1,n+2,\cdots,J)]^2}{I_{2n}(\hat{\theta}_0)}, \tag{20}$$

其中，$\nabla(\hat{\theta}_0;X_j,j=1,2,\cdots,n)$ 和 $\nabla(\hat{\theta}_0;X_j,j=n+1,n+2,\cdots,J)$ 分别指作答序列 1 和 2 在 $\theta=\hat{\theta}_0$ 处对数似然函数的一阶导数。$\nabla(\hat{\theta}_0;X_j,j=1,2,\cdots,n)$ 的展开式详见 Baker 和 Kim (2004, pp. 64–71)。

类似地，

$$S_{max}=\max_{1\leqslant n\leqslant J-1}S_n. \tag{21}$$

S_{max} 检验的虚无假设为：此序列不存在变点；而备择假设为：序列至少存在一处变点。这里也可将 n 的取值范围限定在 J_1 到 $J-J_1$ 间，以增强检测效力。

S_{max} 与 L_{max} 一样，更适用于双侧检验。在单侧检验中，统计量应变更为如下形式：

$$S_{sn}=\begin{cases}\sqrt{S_n}, 若\hat{\theta}_s\geqslant\hat{\theta}_{\bar{s}},\\ -\sqrt{S_n}, 若\hat{\theta}_s<\hat{\theta}_{\bar{s}}.\end{cases} \tag{22}$$

S_{sn} 是单侧检验统计量，各符号意义与 4.1.1 中单侧统计量 L_{sn} 一致。

因此，单侧 PFS 指标 S_s 可表达为：

$$S_s=\begin{cases}=\max\limits_{1\leqslant n\leqslant J-1}S_{sn}, H_0: 对任意n都有\theta_s\geqslant\theta_{\bar{s}},\\ =\min\limits_{1\leqslant n\leqslant J-1}S_{sn}, H_0: 对任意n都有\theta_s\leqslant\theta_{\bar{s}}.\end{cases} \tag{23}$$

使用方法与前述 L_s 一致。

4.1.4 基于加权残差(weighted residual)的 PFS

为探测被试在心理测验中的后期随机作答(back random responding)行为，Yu 和 Cheng (2019)构建了基于加权残差的 PFS 指标。对于测验中正常反应的被试而言，其观察得分模式会与期望得分模式十分接近。而对于异常反应被试，观察与期望得分模式之间会产生较大的偏离。基于加权残差 PFS 的原理在于：找到某个能够将完整作答序列划分为两个子序列的点，该点可使两个子序列的平均绝对加权残差(ABWR; average absolute weighted residual)之间的差值最大化。具体构造流程如下：

Yu 和 Cheng (2019)的研究基于多级计分的心理测验，加权残差 $r_j(\hat{\theta})$ 公式为：

$$r_j(\hat{\theta})=\frac{X_j-E(X_j\mid\hat{\theta})}{P(X_j\mid\hat{\theta})}, \tag{24}$$

式中分子即观察与期望得分间残差的表达式，表示观察与期望得分之间的偏离程度。分母是对于给定能力为 $\hat{\theta}$ 的被试，他在第 j 题上的得分为 X_j 的概率。在 0-1 计分下，加权残差可以表示为：

$$r_j(\hat{\theta})=\frac{X_j-P(X_j=1\mid\hat{\theta})}{P(X_j\mid\hat{\theta})}. \tag{25}$$

进一步地，

$$R_n=\frac{1}{J-n}\sum_{j=n+1}^{J}|r_j(\hat{\theta}_{1n})|-\frac{1}{n}\sum_{j=1}^{n}|r_j(\hat{\theta}_{1n})|, \tag{26}$$

由于 Yu 和 Cheng (2019)侦测的是后期随机作答现象，因此这里只使用到由变点前的正常作答序列计算的能力值 $\hat{\theta}_{1n}$。而且，不局限于后期随机作答，只要是侦查在测验后期出现的异常情况，均可采用(26)式。若要侦查测验前期异常，那么式子可以转变为：

$$R_n=\frac{1}{n}\sum_{j=1}^{n}|r_j(\hat{\theta}_{2n})|-\frac{1}{J-n}\sum_{j=n+1}^{J}|r_j(\hat{\theta}_{2n})|, \tag{27}$$

最终，基于加权残差的 PFS，即 R_{max} 的公式为：

$$R_{max}=\max_{1\leqslant n\leqslant J-1}R_n. \tag{28}$$

与前述 L_{max}、W_{max} 和 S_{max} 三种指标不同的是，R_{max} 在 n 给定下的统计量 R_n 不用于检验虚无假设 $\theta_{1n}=\theta_{2n}$，而是检验测验前期(后期)是否发生异常反应。对于 R_n 而言，某子序列的 ABWR 反映了该子序列观察与期望得分模式之间的偏离程度，当变点前后子序列 ABWR 的差值超过了一定范围，便可说明该被试在测验前期(后期)出现了异常反应。与其它指标相比，R_{max} 更适合用在低风险的心理测验之中。心理测验中常由于被试作答动机缺失导致随机作答产生。然而，随机作答不一定意味被试特质水平的变动，本身持中立观点的被试(即 θ 值在 0 附近)在随机作答的情况下能力估计值可能不会发生改变。假设 R_{max} 的临界值为某正数 h，那么异常反应的判定标准为：如果 R_{max} 显著大于 h，对于(26)式而言，说明被试在测验后期出现了异常反应，而对于(27)式，则说明前期出现异常。

4.1.5 CPA 四种常用的 PFS 简评

CPA 的四种 PFS 的基本原理都在于：判断是否存在可将被试作答序列划分为统计学属性上具有根本差异两部分的点，并定位该点位置。L_{max}、W_{max}、S_{max} 和 R_{max} 四种指标具有不同特性，L_{max}、W_{max} 和 S_{max} 的统计量 L_n、W_n 和 S_n 用于检验虚无假设 $\theta_{1n}=\theta_{2n}$，因此 L_{max}、W_{max} 和 S_{max} 作为双侧检验指标而存在。即当侦查目标仅是检测序列是否存在异常反应，未对异常类型有明确限定时，此时使用这些指标是比较好的。当然，这三种指标也有单侧形式，当目标是侦查具体的异常反应类型(如加速作答)时，适合使用单侧指标。并且，在具体应用层面上，L_{max}、W_{max} 和 S_{max} 三种指标更适用于高风险(high-stakes)、大规模(large-scale)的教育测验。而 R_{max} 因其本身特性，它的统计量 R_n 不用于检验虚无假设 $\theta_{1n}=\theta_{2n}$，而是检验测验前期或后期是否存在异常。当测验管理者对于待侦测的目标效应有明确了解时，例如已明确了侦测目标是后期随机作答，此时使用 R_{max} 是合适的。在应用层面上，R_{max} 更适用在低风险(low-stakes)的心理测验当中。

Sinharay (2016)在计算机自适应测验的环境下对 L_{max}, W_{max} 和 S_{max} 三种 PFS 实施了模拟研究。结果表明：三种 PFS 中，基于 Wald 检验的 W_{max} 效力最高，基于似然比检验的 L_{max} 效力其次，而基于得分检验的 S_{max} 效力最低。Yu 和 Cheng (2019)将 L_{max}, W_{max}, S_{max} 和 R_{max} 四种 PFS 一同用于探测后期随机作答，并对它们的探测性能进行比较。结果发现：四种 PFS 对于侦测任务的一型错误率(Type-I error rate)都控制得很好，但是 R_{max} 的效力要比其他三种 PFS 高出 17%到 42%。

4.2 CPA 中 PFS 临界值的确定方法

在使用 CPA 进行异常反应侦测时，必须借助 PFS，因此 PFS 临界值的确定十分重要。如果 PFS 的临界值选取得不合适，侦测的准确性会大幅降低，导致 CPA 的价值大打折扣。目前对于 L_{max}, W_{max}, S_{max} 和 R_{max} 临界值 L_C, W_C, S_C 和 R_C 的获取，研究者们提出了多种方法。在此介绍两种使用较广的方法：Worsley (1979)提供的蒙特卡罗模拟(Monte Carlo simulation)的方法以及 Storey 和 Tibshirani (2003)提出的 FDR 控制的方法。

4.2.1 蒙特卡罗模拟

此方法的具体步骤如下：

1)模拟 10000 名被试的作答，被试的能力分布从 N (0, 1)中抽取，通过能力参数和已知的题目参数生成这些被试的作答矩阵，因此这些被试都视为正常作答，这一步共重复 B (如 B = 50)次。

2)根据每次重复下每名被试的作答数据可以计算出其 L_{max}, W_{max}, S_{max} 或 R_{max}。注意：此处计算中使用的是估计能力值 $\theta_{Estimate}$。

3)每一次重复下的 10000 个 L_{max}, W_{max}, S_{max} 或 R_{max} 构成虚无假设分布，若设定显著性水平为 0.05，则当中第 500 大的数视为相应临界值的估计值，然后将总共 B 次重复条件下所有估计值的平均数作为 L_C、W_C、S_C 和 R_C 的最终取值。

4.2.2 FDR 控制法

在一项包含 N 名被试的测验中，需要同时检验 N 个假设(即对每名被试是否作答异常进行检验)，需要比较 N 次 PFS 与临界值的大小，这属于多重比较(multiple test)。Shao 等人(2016)认为，此时临界值的设定不能按照普通做法以 0.05 或 0.01 为显著性水平，而应进行校正。一般有两种常用校正方式，一种是 Bonferroni 校正(Bonferroni correction)，将显著性水平校正为 0.05/N 或者 0.01/N，但由于实际测验中样本容量 N 的值很大，所以这种方法过于严格保守，在每一次假设检验中都很难拒绝虚无假设；另外一种方法是控制错误发现率(false discovery rate, FDR; Benjamini & Hochberg, 1995)。在基因学研究中这种方法经常用于多重比较的校正(Benjamini & Hochberg, 1995; Genovese, Lazar, & Nichols, 2002; Li, Witten, Johnstone, & Tibshirani, 2012; Schwartzman & Lin, 2011)。FDR 表示错误标记的数目占标记总数的期望比例，此方法核心思想在于将错误发现率控制在可接受的水平。例如在一次测验中，经侦查后将 100 名被试标记为异常作答，在这 100 名被试中，90 名是真正异常作答的个体，另外 10 名其实是正常作答的个体，属于错误标记，因此这里的 FDR 值为 0.1。在此介绍 Storey 和 Tibshirani (2003)提供的步骤，为便于讲解，仅以 L_C 的确定过程为例进行阐述，具体步骤如下：

1)在已有实测数据的情况下，重新排列每名被试作答数据的顺序，然后计算单次排序下所有人的 L_{max}，总共重排 B 次。每次重排序后所有被试的 L_{max} 集合视作虚无假设分布。

2) T 为临界点 L_c 的取值，是未知数。在此有

如下公式：

$$FDR = \frac{B^{-1}\sum_{b=1}^{B}\sum_{i=1}^{N} I(L_{\max} > T)}{\sum_{i=1}^{N} I(L_{\max} > T)}, \quad (29)$$

其中，b 是重排序的序号，I 为指示函数，当 $L_{max} > T$ 时，$I = 1$，即因指标超过临界值被标记为异常，否则，$I = 0$。FDR 取值可以根据研究和应用的需要人为设定，统计学界一般建议设置为 0.2。于是可解得一个最小的 T 值满足 $FDR \leqslant 0.2$，如此便得到了临界点 L_C 的值。

FDR 控制法的原理在于：分母是被标记为异常反应的被试总数，分子是对于总共 B 次重排序而言，被标记为异常反应的被试的平均数。每次重排序后所有被试的作答序列都视作正常作答序列，故每次重排序下所有 PFS 构成虚无假设分布，分布中大于临界值的 L_{max} 都认为是错误标记。因此，公式(29)充分解释了“FDR 是错误标记的数目占标记总数的期望比例”这一定义。

4.2.3 CPA 中 PFS 临界值的确定方法简评

蒙特卡罗模拟与 FDR 控制法各有所长。蒙特卡罗模拟通过生成的被试参数 θ 和已知的题目参数产生模拟作答，然后在一定显著性水平下取顶端数值的平均数作为临界值，这种方法较为简便易行，但显得比较粗糙。而 FDR 控制法考虑到了多重比较中显著性水平的校正，这是一个实际的、需要重视的问题。这种方法控制了异常反应侦测中的错误发现率，使错误发现率处于可接受的水平，这符合实际情况。因此，FDR 控制法更适合应用于心理与教育测量领域。

5 CUSUM 与 CPA 的综合分析与比较

5.1 CUSUM 与 CPA 基本思路的分析与比较

CUSUM 与 CPA 同属于异常反应侦测的方法，用于分析被试作答序列中是否存在转变点，从更为广泛的层面上而言，两者都可纳入“变点分析”的范畴。但是两种方法从基本思路上而言完全不同，CUSUM 按照题目顺序依次将一系列正或负的残差(观察与期望得分间的残差)累加求和，以得到单侧和双侧 PFS，当然，CUSUM 可以累加的并不限于残差，还可以是对数似然比等内容。因此，该方法在每题后都可更新 PFS 的值。而 CPA 的基本思想在于判断被试的整个作答序列是否可以在某点处划分为两个子序列，这两个子序列的某种统计学属性上的差距会足够大，CPA 的 PFS 可量化这种差距，并且精准定位变点位置。

5.2 CUSUM 与 CPA 优缺点的分析与比较

CUSUM 和 CPA 各有千秋，表 1 陈列了这两种方法的各项特性。CUSUM 的最大优势在于它提供了一种可视化的模式，能快速清楚地获知异常反应发生的位置。并且，在 CAT 中测验人员还可使用 CUSUM 实施过程监控，及时地干预被试作答。此外，它还具有一项较大的优势：通过观察图像，CUSUM 可以清晰、直观、便捷地进行多变点(multiple change points)分析。但是，CUSUM 的缺点在于：它必须人工检查输出的图像以定位变点，并且定位准确性相较 CPA 更差。与之对应，CPA 的优点是：它不仅可以判断某被试是否出现异常反应，还能自动精确地定位变点。CPA 无需像 CUSUM 一样通过观察图像来寻找变点，而是

表 1 CUSUM 与 CPA 的综合比较

	CUSUM	CPA
主要思想	按照题目顺序依次将各题上观察与期望得分间的残差累积求和。	找到某个可将序列划分为具有不同统计学属性两部分的点。
PFS	基于题目平均加权残差的单侧指标 C_j^+，C_j^- 和双侧指标 C^T。	双侧指标：基于似然比检验的 $L_{\max}$，基于 Wald 检验的 $W_{\max}$，基于得分检验的 $S_{\max}$ 和基于加权残差的 $R_{\max}$，以及各自的单侧形式。
单双侧指标	在侦测前已明确目标效应时用单侧指标，未明确目标效应或对目标效应不作具体要求时用双侧指标。	
优点	输出图像，可用于过程监控。	自动精确定位变点。
缺点	需人工检查图像来定位变点，准确性较低。	当变点位于序列最前或最后几题时难以定位。
适用情境	变点前后模型参数已知。	变点前后模型参数未知。其中 $L_{\max}$、$W_{\max}$ 和 $S_{\max}$ 适用于高风险(教育)测验，$R_{\max}$ 适用于低风险(心理)测验。

直接通过 PFS 得到“变点存在与否”的结论，如果变点存在，直接定位它最有可能的位置。这在大规模测验中尤其重要，CPA 方法极大地节省了人力资源，这是它的一大优势。然而，CPA 也有缺陷，当变点位于序列最前或最后几题时，CPA 的检测效力将会大受影响，此时很难对异常反应实施侦测，且难以精确定位变点。原因在于过少的题量无法对被试能力值 θ 形成正确估计，因此正如 4.1.2 中介绍的：Andrews (1993)建议将变点探测范围限定在整个作答序列的中间约 70%的范围。然而在实际应用中，变点位置却常位于此范围之外。

Sinharay (2016)在 CAT 的环境下实施研究，发现了基于 CPA 的 PFS (L_{max}、W_{max} 和 S_{max})侦测效力会优于基于 CUSUM 的 PFS。在关于实际应用中应该选用哪种方法的问题上，Hawkins, Qiu 和 Kang (2003)认为：当变点前后被试的作答模型已知的情况下，采用基于 CUSUM 的 PFS 会更加有效；然而若有一个或多个模型参数未知，则基于 CPA 的 PFS 更优。在心理与教育测验中，模型参数难以精确估计，根据包含大量异常反应数据的原始作答矩阵所估计的参数是不够可靠的。因此，在实际检测中，CPA 要优于传统的 CUSUM，所以应当在心理与教育测量学的研究与应用中推广这种新方法。

6　问题与展望

在心理与教育测验普遍重视测验信效度、测验安全的大环境下，异常反应侦查已成为一项不可忽视的课题，并且具有重要的理论与实践意义。当前对异常反应侦查的研究需求已十分紧迫，亟需得到开展和深入。异常反应现象在测验中十分常见，会直接影响研究结论的可靠性与可推广程度。Shao 等人(2016)认为，数据分析时若存在加速作答数据会使题目和被试参数估计产生偏差，而有偏参数会导致测验管理者对分数作出错误解释进而实施不正确决策。因此，当下亟需开发并完善行之有效的异常反应侦查方法。CPA 在医学、气象、经济等领域的应用已有很长一段时间，但是在近些年才开始“移植”到心理与教育学领域。它相较于传统方法具有较大的优势，能够自动精确地检测出变点位置，有效节省人力资源。CPA 结合了新一代测量理论——项目反应理论，将之运用于心理与教育测量领域，可为测验人员提供极大的便利，帮助高效准确地甄别异常反应被试并对作答数据进行清洗以提高参数估计精度。

虽然 CPA 在异常反应侦查中具有种种优势，但在实际情境的应用中必须注意：绝不可单凭该方法对被试进行分类。CPA 归根到底只是一种统计学方法，它对于被试的分类——即“是否存在异常作答”——只是一种统计学推论，只能作为一种鉴别异常反应被试的辅助手段。除了 CPA 以外，还需要其他来源的证据支持，如座位次序图表、视频监控、教师评价等信息，才可以将某人真正界定为异常反应。本文介绍的 CUSUM 与 CPA 一样同属统计学方法的范畴，因此同理。这一点在教育测验中尤为重要：仅通过 CPA 方法就将某人视为作弊者，进而对其作出处理，这种简易的论断是既不合理也不应该的。正如 2013 年美国教育部(United States Department of Education, 2013)所指出的：统计分析是推论性的，不能仅凭此下最终定论。因此，必须要正视 CPA 的局限性，不可以过度使用此方法，它的价值更多地在于通过清洗异常数据来提高参数估计的精度，从而使研究结论更具可靠性。例如，Shao (2016)依托 CPA 算法设法了一种迭代程序来修正加速作答影响下的参数估计：首先使用原始数据估计参数，并将参数估计的结果用在 CPA 中侦测加速作答被试，然后移除加速部分序列，再使用清理后的数据重新估计参数，上述步骤反复进行，直至满足终止规则。研究结果表明：这项迭代程序可以较大地提高参数估计精度。并且，通过侦测出各被试的具体变点位置，测验管理者可据此修正测验的题数和时长，以减少受时间压力影响在测验后期出现加速作答的被试人数(Shao et al., 2016)。当前 CPA 在心理与教育学界比较“新”，且国内心理与教育测量领域内对 CPA 的研究还处于一片空白，因此未来的研究方向较为广阔。现对 CPA 研究中存在的一些问题及未来可能的研究方向提供一些建议，供后续研究者参考。

6.1　多变点情况下异常反应侦查

本文只讨论了作答序列中存在一处变点的情况，未对多变点分析进行介绍。事实上，当前在心理与教育测量学界，多变点分析的研究还相当少，但现实中多变点现象时常出现，实际测验中可能出现两种或多种效应出现在同一名被试作答过程的现象，如某被试在测验初期存在练习效应，中

期存在疲劳效应，后期存在加速作答。如此一来个体内能力水平可能会发生数次变化，作答序列会存在多个变点。在心理与教育测量之外的领域，多变点分析的常用方法是二值分割法(binary segmentation, BS; Vostrikova, 1981)：首先在一个完整序列中找出某个最可能的变点，它将该序列划分为两个子序列，然后在这两个子序列中继续寻找变点，将子序列划分为更小的序列，此步骤不断循环，直到满足标准后终止，如此一来便找到了多个变点。BS 可以很方便迁移到心理与教育测量学中，以深化多变点分析侦测异常反应的研究。在现今的测量学研究与应用中，多变点比单变点可能占据更重要的地位，某种意义上而言多变点分析的研究具有更大的意义。今后应着眼于多变点异常反应的 IRT 模型构建以及指标和方法开发的一系列研究。

6.2　结合反应时的异常反应侦查

当前对于异常作答的侦查主要根据被试在各题上的得分数据，然而仅凭此类数据会产生较多的判断失误。因此，有研究者建议异常反应侦查时可以结合其他方面的信息来增强检测效力，例如充分利用座位次序图表，视频监控和后续面谈(Tendeiro & Meijer, 2014)等信息。并且，就目前而言，反应时是一种较容易获得且十分有效的信息，基于计算机的测验可以很好地收集被试在各题上的反应时。目前为止，研究者开发了一系列反应与反应时联合建模的模型，包括四参数 logistic 反应时模型(four-parameter logistic response time model, 4PL-RTM; Wang & Hanson, 2005)和层级框架模型(hierarchical framework model; van der Linden, 2007; Fox & Marianti, 2016)等，这些模型足以支持 CPA 的研究。Wang 和 Xu (2015)结合被试作答反应和反应时数据建立了混合层级模型(mixture hierarchical model)，对快速猜测行为(rapid guessing behaviour，即由时间压力或动机缺失导致的随机作答行为)实施侦测，取得了比较好的检测效果。此外，存在加速作答的被试在测验后期各题上作答的反应时会更短(Shao et al., 2016)。因此，Shao (2016)在研究中使用基于反应时数据的 CPA 对加速作答行为实施侦测，结果发现：在侦测结果中不仅一型错误率得到了良好控制，方法的效力也很高。这说明了仅通过反应时数据实施侦查也可以取得不错的效果。通过结合上述信息，可极大提升 CPA 的侦查效力。因此，在实际应用中可以考虑结合其他来源的有效信息来提高侦查的准确性，这是一个有价值的研究方向。

6.3　基于非参数化 PFS 的异常反应侦查

当前研究者已经开发出了四类 CPA 的 PFS 指标：基于似然比检验的 L_{max}、基于 Wald 检验的 W_{max}、基于得分检验的 S_{max} 和基于加权残差的 R_{max}。四类 PFS 都基于 IRT 构建，同属参数化指标的范畴。在 CPA 领域，目前尚无关于非参数化指标的研究。非参数化相较于参数化的方法更具简洁性，而且某些情况下非参数化会比参数化指标表现得更好。例如, Karabatsos 和 George (2003)对 36 种传统 PFS 指标进行比较研究后发现：非参数化的 PFS 比参数化的侦查效力更高。原因可能在于：计算参数化的 PFS 过程中需要对同一数据集使用两次——第一次用于估计 IRT 参数，第二次则运用这些参数对数据进行拟合，即计算 PFS。因此参数会和数据产生关联，而非参数化的 PFS 则没有这种关系。因此，在将来 CPA 的研究中，可对非参数化 PFS 构建这一方向多加考虑。

6.4　多级评分以及多维测验下的异常反应侦查

在如今心理测量学领域，多级计分的量表占据了主体地位。但是本文中列举的各项 PFS 中，只有 Yu 和 Cheng (2019)提出的基于加权残差的 R_{max} 是建立在多级计分测验上的，而且向两级计分的指标转化也很方便。因此，可以考虑将现有的 PFS 拓展至多级计分，以增加这些指标的适用范围，此类研究难度较低，可行性较高(Sinharay, 2016)，并且具有较大的应用价值。除了将指标拓展至多级计分以外，也可将现有的 PFS 向多维测验进行拓展，多维量表的开发是当前趋势所在，例如在基于英文语言的数学测验中，每道题上同时考察英语与数学两个维度的能力，如果某考生存在加速作答，那么经过变点之后其英语和数学能力都将降低。当前多维项目反应理论(multidimensional IRT, MIRT)已较为成熟，可以支撑多维异常反应侦查研究的开展。因此，这也是一项具有可行性和价值的工作。

参考文献

陈希孺. (1991). 变点统计分析简介. *数理统计与管理*, (1), 52–59.

Abahous, H., Ronchail, J., Sifeddine, A., Kenny, L., &

Bouchaou, L. (2018). Trend and change point analyses of annual precipitation in the Souss-Massa Region in Morocco during 1932–2010. *Theoretical and Applied Climatology, 134*(3-4), 1153–1163.

Allen, D. E., McAleer, M., Powell, R. J., & Singh, A. K. (2018). Non-parametric multiple change point analysis of the global financial crisis. *Annals of Financial Economics, 13*(02), 1850008.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems, 51*, 339–367.

Andrews, D. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica, 61*(4), 821–856.

Armstrong, R. D., & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement, 33*(5), 391–410.

Baker, F. B., & Kim, H. S. (2004). *Item response theory: Parameter estimation techniques (2nd ed.)*. New York, NY: Marcel Dekker.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*(4), 331–348.

Bolt, D. M., Mroch, A. A., & Kim, J.-S. (2003, April). *An empirical investigation of the hybrid IRT model for improving item parameter estimation in speeded tests*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Bradlow, E., & Weiss, R. E. (2001). Outlier measures and norming methods for computerized adaptive tests. *Journal of Educational and Behavioral Statistics, 26*(1), 85–104.

Bradlow, E., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93*, 910–919.

Chen, J., & Gupta, A. K. (2012). *Parametric statistical change point analysis: With applications to genetics, medicine, and finance* (2nd ed.). New York: Springer.

Csorgo, M., & Horvath, L. (1997). *Limit theorems in change-point analysis*. New York, NY: Wiley.

de Boeck, P., Cho, S. J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement, 35*(8), 583–603.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum, Inc.

Estrella, A., & Rodrigues, A. (2005). *One-sided test for an unknown breakpoint: Theory, computation, and application to monetary theory (Staff Reports No. 232)*. Federal Reserve Bank of New York.

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement, 9*, 123–131.

Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate behavioral research, 51*(4), 540–553.

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage, 15*(4), 870–878.

Goegebeur, Y., de Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika, 73*(1), 65.

Hawkins, D. M., Qiu, P., & Kang, C. W. (2003). The changepoint model for statistical process control. *Journal of Quality Technology, 35*(4), 355–366.

Hong, M. R., & Cheng, Y. (2019). Robust maximum marginal likelihood (RMML) estimation for item response theory models. *Behavior Research Methods, 51*(2), 573–588.

Karabatsos, & George. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277–298.

Kass-Hout, T. A., Xu, Z., McMurray, P., Park, S., Buckeridge, D. L., Brownstein, J. S., ... Groseclose, S. L. (2012). Application of change point analysis to daily influenza-like illness emergency department visits. *Journal of the American Medical Informatics Association, 19*(6), 1075–1081.

Lai, T. L. (2001). Sequential analysis: Some classical problems and new challenges. *Statistica Sinica, 11*(2), 303–408.

Lee, Y. H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika, 78*(3), 557–575.

Li, J., Witten, D.M., Johnstone, I.M., & Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation

for RNA-sequencing data. *Biostatistics, 13*(3), 523–538.

Maleki, S., Bingham, C., & Zhang, Y. (2016). Development and realization of changepoint analysis for the detection of emerging faults on industrial systems. *IEEE Transactions on Industrial Informatics, 12*(3), 1180–1187.

Meade, A. W. (2016). Understanding and detecting careless responding in survey research. Retrieved February 15, 2020, from https://cba.unl.edu/outreach/carma/documents/CARMA-Meade-Presentation.pdf

Meijer, R. R. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement, 39*(3), 219–233.

Mortaji, S. T. H., Noorossana, R., & Bagherpour, M. (2015). Project completion time and cost prediction using change point analysis. *Journal of Management in Engineering, 31*(5), 04014086.

Nam, C. F. H., Aston, J. A. D., & Johansen, A. M. (2012). Quantifying the uncertainty in change points. *Journal of Time Series Analysis, 33*(5), 807–823.

Nigro, M. B., Pakzad, S. N., & Dorvash, S. (2014). Localized structural damage detection: A change point analysis. *Computer-Aided Civil and Infrastructure Engineering, 29*(6), 416–432.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*(3), 200–219.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika, 41*(1-2), 100–115.

Patton, J. M., Cheng, Y., Hong, M. R., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics, 44*(3), 309–341.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271–282.

Schwartzman, A., & Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika, 98*(1), 199–214.

Shao, C. (2016). *Aberrant response detection using change-point analysis* (Unpublished Doctoral dissertation). University of Notre Dame, Notre Dame, IN.

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrica, 81*(4), 1118–1141.

Sinharay, S. (2016). Person fit analysis in computerized adaptive testing using tests for a change point. *Journal of Educational and Behavioral Statistics, 41*(5), 521–549.

Sinharay, S. (2017a). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics, 42*(1), 46–68.

Sinharay, S. (2017b). Some remarks on applications of tests for detecting a change point to psychometric problems. *Psychometrika, 82*(4), 1149–1161.

Sinharay, S. (2017c). Which statistic should be used to detect item preknowledge when the set of compromised items is known?. *Applied Psychological Measurement, 41*(6), 403–421.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences, 100*(16), 9440–9445.

Suh, Y., Cho, S. J., & Wollack, J. A. (2012). A comparison of item calibration procedures in the presence of test speededness. *Journal of Educational Measurement, 49*(3), 285–311.

Suhaila, J., & Yusop, Z. (2018). Trend analysis and change point detection of annual and seasonal temperature series in Peninsular Malaysia. *Meteorology and Atmospheric Physics, 130*(5), 565–581.

Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement, 36*(5), 420–442.

Tendeiro, J. N., & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement, 51*(3), 239–259.

Thies, S., & Molnár, P. (2018). Bayesian change point analysis of Bitcoin returns. *Finance Research Letters, 27*, 223–227.

United States Department of Education. (2013). *Testing integrity: Issues and recommendations for best practice*. Retrieved November 21, 2019, from http://nces.ed.gov/pubs2013/2013454.pdf.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287–308.

van Krimpen-Stoop, E. M. L. A., Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & G. A. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 201–219). Dordrecht, Netherlands: Springer.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*(2), 199–217.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement, 26*(2), 164–180.

Vostrikova, L. Y. (1981). Detecting “disorder” in multidimensional random processes. *Doklady Akademii*

Nauk, 259(2), 270–274.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456–477.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29*(5), 323–339.

Wollack, J. A., & Cohen, A. S. (2004, April). *A model for simulating speeded test data*. Paper presented at the meeting of the American Educational Research Association. San Diego, CA.

Worsley, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association, 74*, 365–367.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). New York: Waxmann.

Ye, W., Liu, X., & Miao, B. (2012). Measuring the subprime crisis contagion: Evidence of change point analysis of copula functions. *European Journal of Operational Research, 222*(1), 96–103.

Yu, M., & Ruggieri, E. (2019). Change point analysis of global temperature records. *International Journal of Climatology, 39*(8), 3679–3688.

Yu, X., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods, 24*(5), 658–674.

Zhang, J. (2014). A sequential procedure for detecting compromised items in the item pool of a CAT system. *Applied Psychological Measurement, 38*(2), 87–104.

Change point analysis: A new method to detect aberrant responses in psychological and educational testing

ZHANG Longfei; WANG Xiaowen; CAI Yan; TU Dongbo

(*School of Psychology, Jiangxi Normal University, Nanchang* 330022, *China*)

Abstract: The change point analysis (CPA), as one of the most widely used methods for statistical process control, is introduced to psychological and educational measurement for detection of aberrant response patterns in recent years. CPA outperforms the traditional method as follows: In addition to detecting aberrant response patterns, it can also pinpoint the locations of change points, contributing to efficient cleansing of response data. The method is employed to determine whether there is a point so that the complete sequence can be divided into two parts with different statistical properties, where person-fit statistics (PFS) is needed for quantifying the difference between two sub-sequences. Future researchers should pay more attention to multiple change points detection, making full use of other effective information like response time data, developing non-parametric indices as well as reforming the exiting person-fit statistics for polytomous and multidimensional tests, so as to enhance its applicability and power.

Key words: aberrant responses; change point analysis; cumulative summation; person-fit statistics